Test Integrity – Data Forensics

OSSE's

Next Generation Assessment Meeting

July 28, 2016

Dennis Maynes, Chief Scientist
Caveon, LLC

caveon™
Test Security

Outline

1. Statistical Methodology
2. Statistics
   1. Gains/Losses
   2. Similarity
   3. Answer Changes
3. Summary

caveon™ Test Security

Purpose of Statistical Methodology

- Measure and monitor test security threats.
- Detect anomalies in schools, classrooms and students using test response data.
- Find *potential* misbehavior and test security violations.
- Help learn where, when, by whom, and effects of suspect activity.

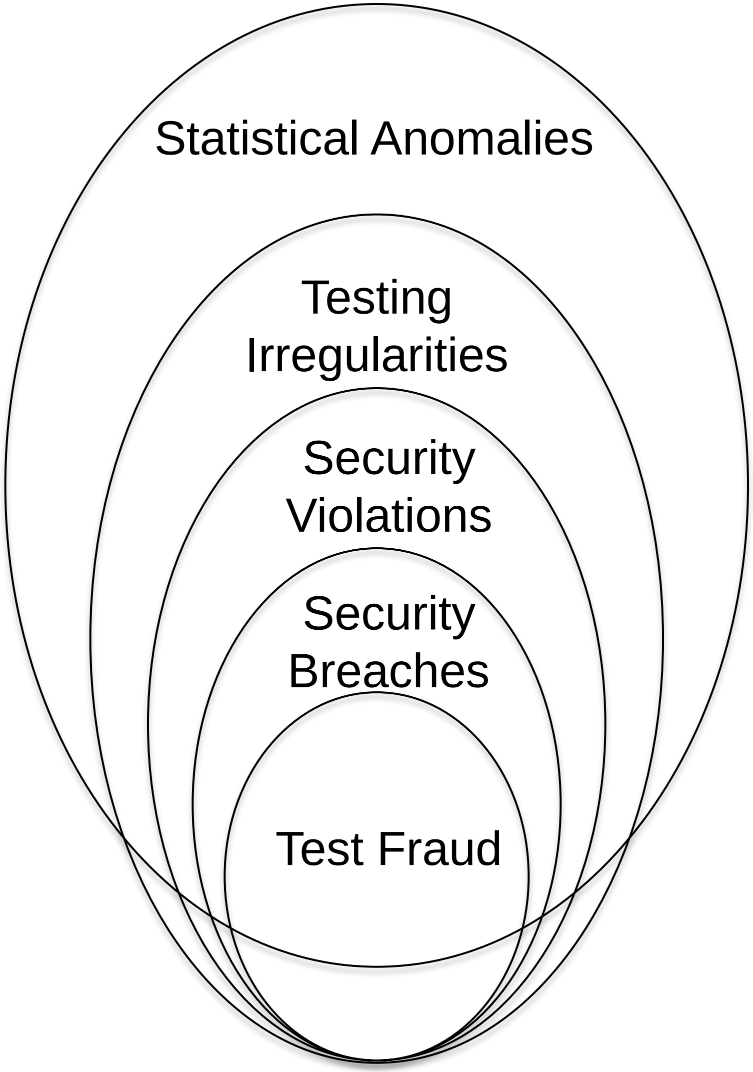
caveon™
Test Security

Definitions

- *Statistical anomalies* are observed data that do not conform to statistical models of normal test taking.
- *Testing irregularities* are abnormal occurrences which may have impacted the test administration.
- *Test security violations* occur when the security protocols of the test have not been followed.
- A *breach in test security* is an event which has jeopardized the fairness and the validity of the current or future test administrations.
- *Test fraud* involves intent by a perpetrator to breach the security of the test.

caveon™ Test Security

Data Forensics Detection


Statistical Anomalies

Testing Irregularities

Security Violations

Security Breaches

Test Fraud

caveon™ Test Security

Small Probabilities → Flags

- Rare and unusual events are improbable.
  - Being struck by lightning or a meteor
  - Winning the lottery again
- Report small probabilities in three ways.
  - Scientific: 1.0e-8 = 0.00000001,
  - Odds: One chance in 100 million (8 zeros), &
  - Index: 8 (count zeros don't print them).
- $p = 10^{-index}$
- Small probabilities identify *potential* test security violations.

Anomalies

- Anomalies provide circumstantial evidence.
- Multiple anomalies are less likely to have occurred through some happenstance than a single anomaly.
- The ultimate goal is to strengthen test security.

Circumstantial evidence is evidence that relies on an inference to connect it to a conclusion of fact.

Most evidence (e.g., finger prints) is circumstantial.

caveon™
Test Security

Patterns: Gains/Losses

Gains analysis will begin with the 2016 PARCC administration.

- Cohorts are computed when the same students are used in year-to-year differences.
- Cross-sections are computed when the same grades used in year-to-year differences.
- Three patterns should be considered.
  - Score increases followed by score decreases
  - Score increases from prior years
  - Score decreases from prior years

caveon™
Test Security

Statistics: Gains/Losses

- Large gains are often triggers or supporting factors in investigations.
- Large losses after implementation of security measures may also initiate investigations.
- Exam fraud is an attempt to gain an unfair advantage.
- Demonstration that an advantage was gained or attempted is needed to support inferences concerning potential fraud.

caveon™
Test Security

Gains/Losses: Context

- Gains can result from
  - Improved teaching
  - Population changes
    - Examples: student mobility, boundary changes
  - Coaching or disclosing actual exam content
- Losses can result from opposite factors and
  - Interrupted exam sessions
  - Lack of student motivation

Gains/Losses: Method

- Match student data from year-to-year.
- Predict score differences using prior scores (regression).
- In order, the preferred scores are:
  - Scale scores (equated scores),
  - Standardized MLE (θ) scores,
  - Percentile scores, and
  - Raw scores.

Gains/Losses: Students

- Compute predicted differences.
- Standardize using regression equation.
- Evaluate Z-score for individual students.
- Convert Z-score to an index value, $\alpha=0.00001$.

Gains/Losses: Groups

- Find concentrations of gains/losses.
- Flag students with gains/losses (α=0.05).
- Compute rate of flagged students.
- Compare the rate in the school against the overall flag rate for the state.
- Compute index (probability) for the school.
  - Hypergeometric: Fisher's Exact Test
  - Multiple comparison α= 0.01.

Gains/Losses: Inference

- What might have occurred to explain the score changes?
- Are the data consistent with propositions for or against score manipulation?

Gains/Losses: Follow Up

- Is student knowledge consistent with scores?
- Was student improvement due to some increased capability? Eye glasses? Language proficiency?
- Seek documentation and information that can help explain the anomalies.

Patterns: Similarity

- Improbable agreement of answers exists between two or more test takers.
- Identical incorrect answers provide more evidence of potential wrong-doing than identical correct answers.
- Non-independence is evidence of potential collusion.
  - Seating charts and proximity
  - Answering questions at the same time
  - Communication between test takers

caveon™
Test Security

Statistics: Similarity

- Demonstrate whether tests were taken independently.
- High index values could indicate
  - Answer-copying & collusion
  - Guessing strategies or thoughtless responding
- Lower index values could indicate
  - Coaching within a group
  - Shared crib sheet
  - Studying together
  - Shared misconceptions of content

caveon™
Test Security

Similarity: Context

- Students learn the same way to wrongly answer questions.
- Studying together is a frequent but not credible explanation because all students study together.
- Data errors (a test appears in the database twice) can artificially induce similarity.
- Similarity detects potential fraud.
  - Shared answer key
  - Copy/communicate with each other
  - Receive assistance from an adult

caveon™ Test Security

Similarity: Method

- Compare every student’s response vector with others in the school.
- Evaluate probabilities of matching answers using IRT models.
- Probabilities depend upon performance.
  - Two students with 100% will have identical correct answers (when one answer is correct)
  - Expected agreement decreases with lower scores
  - Statistical power decreases with higher scores

Similarity: Item Response Theory

- Probability of correct answers depends upon performance.
  - $P(\text{correct}_j | \theta) = [1+\exp(-a_j*(\theta - b_j))]^{-1}$
- Probability of matching correct answers computed using independence.
  - $P(\text{both correct}|\theta_i, \theta_j) = p(\text{correct}|\theta_i) \text{ x } p(\text{correct}| \theta_j)$
- Probabilities for incorrect answers modeled using the Nominal Response Model (NRM).

Similarity: NRM

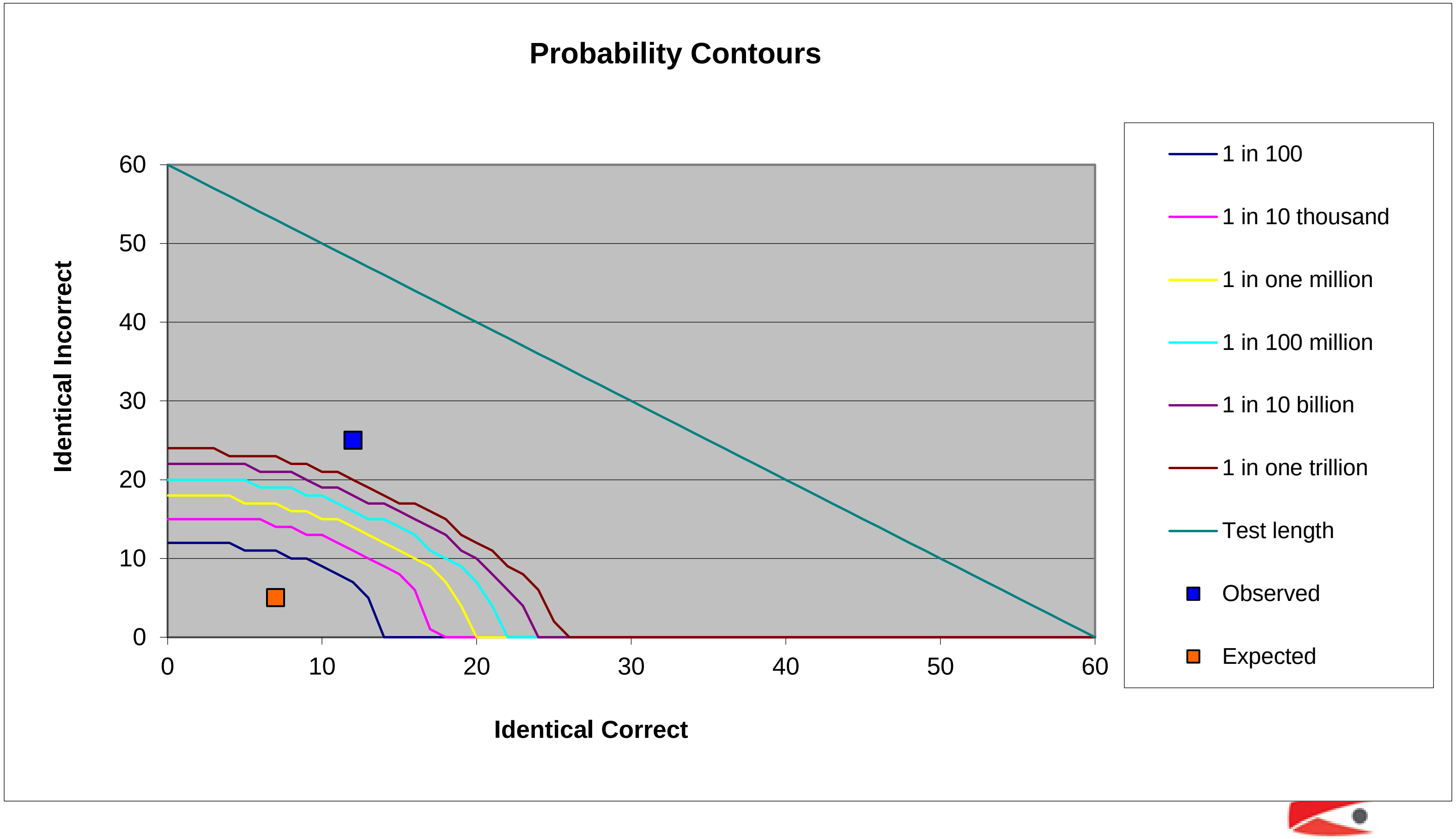
- Bock (1972)

$$p(x_i = k \mid \theta) = \frac{e^{a_{ik}\theta + g_{ik}}}{\sum_{t=1}^{R_i} e^{a_{it}\theta + g_{it}}}$$

- Each response has a probability.
- Probabilities depend upon performance, item difficulty, and item discrimination.
- The model allows computation of match probabilities.

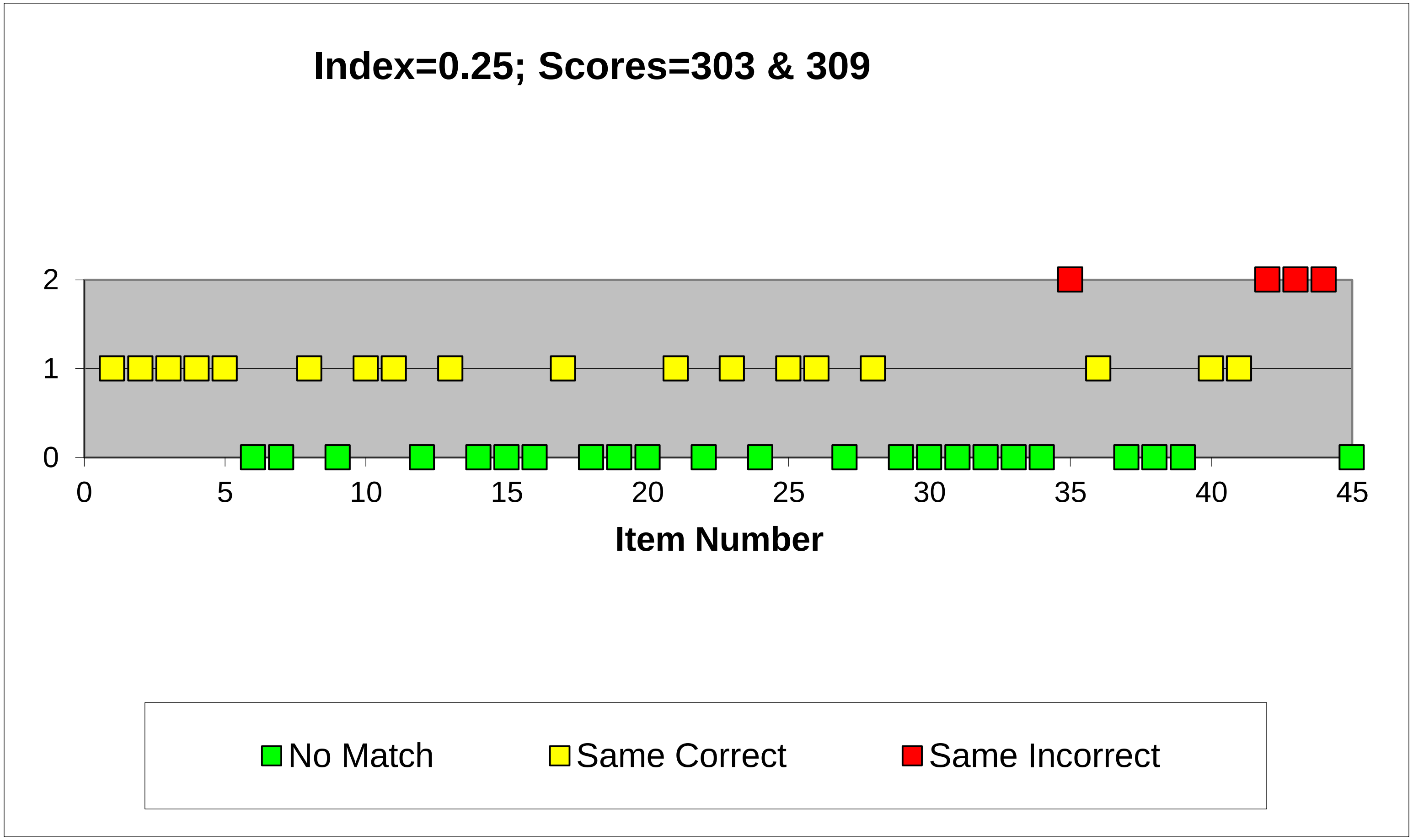
caveon™
Test Security

Similarity: Observed vs. Expected

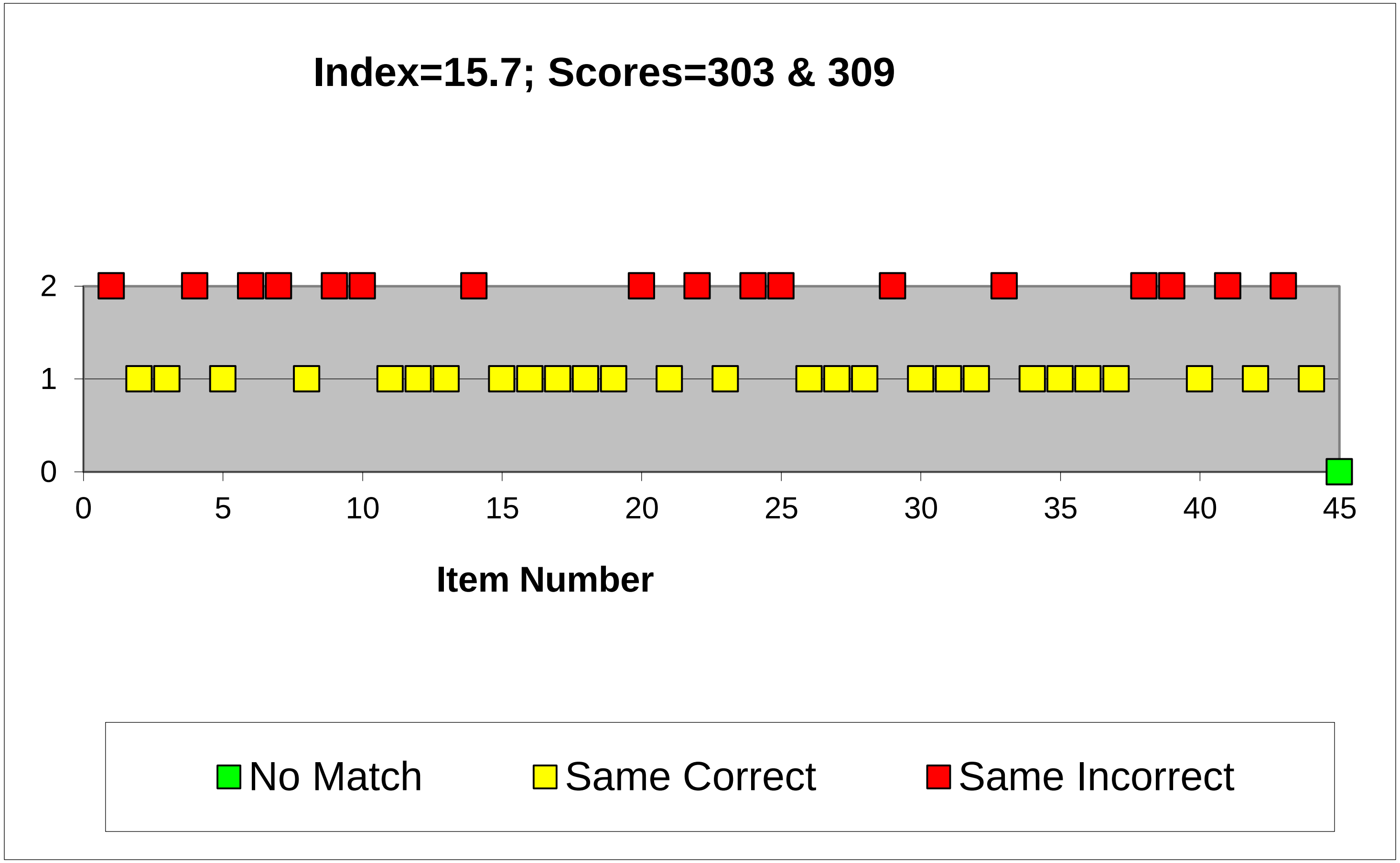


caveon™
Test Security

Similarity: Independence



Similarity: Nonindependence



caveon™
Test Security

Similarity: Students

- Compute observed agreement (identical correct & identical incorrect) for each pair.
- Compute probability using model.
- Adjust for making many comparisons $\alpha=.05/((n\text{-}1)/2)$.
- Flag students when $p < 0.00001$.

Similarity: Clusters

- Clusters can identify groups of students involved in nonindependent test taking.
- Analysis of alignments can help determine whether the similarity includes more than two students.
- Clusters may result from:
  - Communication before/during the test,
  - Coaching by an adult during the test, and
  - Very unusual factors or situations.

Similarity: Groups

- Find concentrations of nonindependence.
- Flag students with similarity (α=0.05).
- Compute rate of flagged students.
- Compare the rate in the school against the overall flag rate for the state.
- Adjust rates for the number of clusters.
- Compute index (probability) for the school.
  - Hypergeometric: Fisher's Exact Test
  - Multiple comparison α= 0.01

Similarity: Inference

- If dependent test taking is to be inferred, it is important to provide a plausible explanation.
  - Is the alignment something that might happen through teaching?
  - How might dependence in responding have occurred?
- Steps are needed to explain what might have happened.
- Are the data consistent with propositions for or against dependence?

Similarity: Follow Up

- Were students allowed to communicate?
- Do flagged students have associations?
- Could test content have been coached?
- What test-taking strategies were taught?
- Seek documentation and information that can help explain the anomalies.

Patterns: Answer Changing

- WTR answer changes increase scores.
- RTW answer changes decrease scores.
- The difference between WTR and RTW is a measure of score change due to answer changing.

Statistics: Answer Changes

- “Erasure” analysis – paper-and-pencil
- Computer records visits, item reviews, and answer changes
  - Analysis depends on what has been recorded
- Potential directions to change answers
- Potential communication to change answers (e.g., while in restroom or searching internet in restroom)

caveon™
Test Security

Answer Changes: Context

- Reviewing & rethinking answers
- Correction of shift errors on paper, and
- Student input behavior on computer
- Marking/eliminating on paper (usually looks different than answer changing)
- Answer copying
- Redirecting & tampering

Answer Changes: Method

- Assume answer changing is sporadic.
- Compute frequencies per item (or common).
  - WTR, RTW, WTW, no changes
- For each test instance:
  - compute probability of observed WTR count (binomial),
  - compute probability of WTR minus RTW difference (trinomial), and
  - convert probabilities into index values.

Answer Changes: Students

- Flag students for high WTR and high WTR-RTW difference when $p < 0.00001$.
- Report # WTR’s and WTR-RTW difference.

Answer Changes: Groups

- Each student contributes one to total, not each changed answer (aka averages).
- Flag individual students (α=0.05).
- Compute rate of flagged students.
- Compare the rate in the school against the overall flag rate for the state.
- Compute index (probability) for the school.
  - Hypergeometric: Fisher's Exact Test
  - Multiple comparison α= 0.01

caveon™
Test Security

Answer Changes: Inference

- If tampering is to be inferred, it is important to provide a plausible explanation.
  - Coaching by an adult: “Check your work”
  - Conversation in restroom followed by answer changing
  - Adult reviewing test session after-the-fact
- Are the data consistent with propositions for or against tampering?

caveon™
Test Security

Answer Changes: Follow Up

- Would students answer questions in the same way?
- Are there patterns in items with WTR and RTW answer changes?
- Is there an association between WTR and student scores?
- Seek documentation and information that can help explain the anomalies.

caveon™
Test Security

Statistics: Other Information

- Identical tests
- Differences between scored and non-scored items
- Special situations which may add clarity:
  - Accommodation
  - Stayers & leavers

caveon™
Test Security

Summary

- Because conservative thresholds are used detection of anomalies is not “automatic.”
- Anomalies are indications of *potential* test security violations, not proof.
- Additional information should be sought.
- Patterns and multiple statistics provide clarity.
- Inferences about scores require information about scores; the same is true for behavior.

caveon™
Test Security

Questions

caveon™

Test Security

Thank You!

Follow Caveon on twitter @caveon

Check out our blog…www.caveon.com/blog

LinkedIn Group – “Caveon Test Security”

caveon™
Test Security